

A Benchmark Suite for Template Detection and Content Extraction^{*}

Julián Alarte¹, David Insa¹, Josep Silva¹, and Salvador Tamarit²

¹ Universitat Politècnica de València, Camino de Vera S/N, E-46022 Valencia, Spain.
{jalarte,dinsa,jsilva}@dsic.upv.es

² Babel Research Group, Universidad Politécnica de Madrid, Madrid, Spain
stamarit@babel.ls.fi.upm.es

Abstract. Template detection and content extraction are two of the main areas of information retrieval applied to the Web. They perform different analyses over the structure and content of webpages to extract some part of the document. However, their objective is different. While template detection identifies the template of a webpage (usually comparing with other webpages of the same website), content extraction identifies the main content of the webpage discarding the other part. Therefore, they are somehow complementary, because the main content is not part of the template. It has been measured that templates represent between 40% and 50% of data on the Web. Therefore, identifying templates is essential for indexing tasks because templates usually contain irrelevant information such as advertisements, menus and banners. Processing and storing this information is likely to lead to a waste of resources (storage space, bandwidth, etc.). Similarly, identifying the main content is essential for many information retrieval tasks. In this paper, we present a benchmark suite to test different approaches for template detection and content extraction. The suite is public, and it contains real heterogeneous webpages that have been labelled so that different techniques can be suitable (and automatically) compared.

1 Introduction

Template extraction is an important tool for website developers, and also for website analyzers such as crawlers. Content extraction is essential for many information processing tasks applied to webpages. In the last decade, there have been important advances that produced several techniques for both disciplines.

^{*} This work has been partially supported by the EU (FEDER) and the Spanish *Ministerio de Economía y Competitividad (Secretaría de Estado de Investigación, Desarrollo e Innovación)* under grant TIN2013-44742-C4-1-R and by the *Generalitat Valenciana* under grant PROMETEO/2011/052. David Insa was partially supported by the Spanish Ministerio de Educación under FPU grant AP2010-4415. Salvador Tamarit was partially supported by research project POLCA, Programming Large Scale Heterogeneous Infrastructures (610686), funded by the European Union, STREP FP7.

Hybrid methods that exploit the strong points of several techniques have been defined too. In order to test, compare and tune these techniques, researchers need:

- collections of benchmarks that are heterogeneous (to ensure generality of the techniques) and
- a gold standard (to ensure the same evaluation criteria).

A benchmark suite is essential to measure the performance of these techniques, and to compare them with previous approaches. Benchmark suites are used in the testing phase and in the evaluation phase. The testing phase allows developers to optimize the techniques by adjusting parameters. Once the technique has been tuned, the evaluation phase allows us to know its performance with objective measures. It is obvious that the set of benchmarks used in the testing phase cannot be used in the evaluation phase, thus, they need disjoint sets of webpages.

In this paper we present a benchmark suite together with a gold standard that can be used for template detection and for content extraction. All benchmarks have been labelled so that every HTML element of the webpages indicates whether it should be classified as main content or not, and whether it should be classified as template or not. The suite also incorporates scripts to automatize the benchmarking process.

This suite has been developed as the result of a research project. We developed a new technique for content extraction [3] that was later adapted for template detection [1]. In the evaluation phase, our initial intention was to use a public benchmark suite. We first tried to use the CleanEval [2] suite of content extraction benchmarks, because it has been widely used in the literature. Unfortunately, it is not prepared for template detection. Then, we contacted the authors of other techniques that had already evaluated their techniques. However, we could not use these benchmarks due to privacy (they belong to a company or project whose results were not shared), copyright (they were not publicly available) or unavailability (they had been lost). Finally, we decided to build our own benchmark suite and make it free and publicly available. The rest of this paper presents that benchmark suite.

2 The TECO Benchmark Suite

TECO (TEmplate detection and COntent extraction benchmarks suite) was created as a benchmark suite specifically designed for template detection and content extraction. It can be used for testing and evaluation of these techniques, and it is formed from 40 real websites downloaded from Internet. We selected heterogeneous websites such as blogs, companies, forums, personal websites, sports websites, newspapers, etc. Some of the websites are well known, like the BBC website or the FIFA website, and others are less known like personal blogs or small companies websites. The downloading of the webpages was done in some

cases using the OS X software SiteSucker, and in other cases using the Linux command `wget`.

It is important to know how the websites were downloaded and stored, so that other researchers can increase the suite if it is needed. The following command downloads a website from the Linux terminal using the `wget` command:

```
$ wget --convert-links --no-clobber --random-wait -r 3 -p -E -e
robots=off -U mozilla http://www.example.org
```

The meaning of the flags used is:

- `--convert-links`: Converts links so they can work locally.
- `--no-clobber`: Do not overwrite any existing file.
- `--random-wait`: Random waits between downloads.
- `-r 3`: Recursive downloading up to 3 levels of links.
- `-p`: Downloads everything.
- `-e robots=off`: Act as not being a robot.
- `-E`: Get the right file extension.
- `-U mozilla`: Identify as a Mozilla browser.

Each benchmark is composed of:

- A principal webpage, called *key page*. It is the target webpage from which the techniques should extract the main content or the template—note that it is not necessarily the main webpage of the website (e.g., `index.html`)—.
- A set of webpages that belong to the same website as the key page. This set contains all those webpages that are linked by the key page, and also the webpages linked by them.

2.1 Producing the gold standard

The suite comes with a gold standard that can be used as a reference to compare different techniques. The gold standard specifies for each key page what parts form the template. This is indicated in the own webpage by using HTML classes that indicate what elements are classified as *notTemplate*. It has been produced manually by careful inspection of the websites and mixing the opinion of several people.

In particular, once all the websites were downloaded (the key page and two levels of linked webpages in the same domain), four different engineers did the following independently:

- They manually explored the key page and the webpages accessible from it to decide what part of the webpage is the template and what part is the main content.
- They printed the template and the main content of the webpage.

Then, the four engineers met and performed again these two actions but now all together sharing their individual opinions. Using the results of this agreement, each website was prepared for both, template extraction and content detection. On the one hand, all elements from the key page not belonging to the template

were included in a HTML class called *notTemplate*. This way, a template extraction tool can automatically compare its output with the nodes not belonging to the *notTemplate* class. On the other hand, all elements belonging to the main content were included in an HTML class called *mainContent*. Therefore, a content extraction tool can easily compare its output with the nodes belonging to that class.

2.2 Benchmark details

A classification of the benchmarks is important and useful depending on the application and technique that is being fed with them. We provide different classifications according to the purpose and properties of the benchmarks. First, all benchmarks have been classified into five groups:

Companies / Shops, Forums / Social, Personal websites / Blogs,
Media / Communication, Institutions / Associations.

Table 1 shows this classification together with the URLs from which we extracted the benchmarks.

Table 2 shows some properties of the benchmarks. Here, column **Nodes** indicates the total number of DOM nodes in the key page, column **T. Nodes** shows the number of DOM nodes that belong to the template and column **M.C. Nodes** refers to the number of DOM nodes that belong to the main content.

The benchmarks were also classified according to the number of webpages that implement the template. Table 3 shows this information. Here, the identifier of the benchmarks (**Id**) comes from Table 2. For each benchmark, column **VL** indicates the number of hyperlinks in the main menu, column **TT** shows the number of webpages accessible from the main menu that implement entirely the template, column **PT** indicates the number of webpages accessible from the main menu that implement partially the template, column **DT** shows the number of pages accessible from the main menu that do not implement the template at all, and finally, column **Notes** explains, when applicable, why not all webpages implement the template.

Table 1: Sources of the benchmarks

Website type	Original URL of the webpage
Companies / Shops	clotheshor.se www.emmaclothes.com www.mediamarkt.es www.ikea.com/gb/en.html www.swimmingpool.com www.skipallars.cat/en.html www.turfparadise.com www.beaches.com www.felicity.co.uk www.us-nails.com www.wayfair.co.uk catalog.atsfurniture.com www.glassesusa.com www.mysmokingshop.co.uk
Forums / Social	stackoverflow.com www.filmaffinity.com/es/main.html
Personal / Blogs	users.dsic.upv.es/~jsilva/wwv2013/index2.html users.dsic.upv.es/~dinsa/en/index.html labakeryshop.com
Media / Communication	www.history.com www.tennis.com www.tennischannel.com riotimesonline.com www.engadget.com www.bbc.co.uk/news www.vidaextra.com en.citizendium.org edition.cnn.com www.lashorasperdidas.com www.thelawyer.com
Institutions / Associations	water.org www.jdi.org.za www.eclipse.org www.landcoalition.org es.fifa.com cordis.europa.eu/fp7/ict/fire.html www.cleanclothes.org www.ox.ac.uk/staff/index.html clinicaltrials.gov/ct2/search/index/index.html www.informatik.uni-trier.de/~ley/pers/hd/s/Silva_Josep.html

Table 2: Benchmark properties

Id	Benchmark	Nodes	T. Nodes	M.C. Nodes
1	water.org/index.html	948	711	237
2	www.jdi.org.za/index.html	626	305	225
3	stackoverflow.com/index.html	6450	447	6003
4	www.eclipse.org/index.html	256	156	100
5	www.history.com/index.html	1246	669	260
6	www.landcoalition.org/index.html	1247	393	588
7	es.fifa.com/index.html	1324	276	737
8	cordis.europa.eu/fp7/ict/fire.html	959	335	179
9	clotheshor.se/index.html	459	231	228
10	www.emmaclothes.com/index.html	1080	374	706
11	www.cleanclothes.org/index.html	1335	266	1069
12	www.mediamarkt.es/index.html	805	337	40
13	www.ikea.com/gb/en.html	1545	407	1138
14	www.swimmingpool.com/index.html	607	499	176
15	www.skipallars.cat/en.html	1466	842	573
16	www.tennis.com/index.html	1300	463	676
17	www.tennischannel.com/index.html	661	303	148
18	www.turfparadise.com/index.html	1057	726	322
19	riotimesonline.com/index.html	2063	879	969
20	www.beaches.com/index.html	1928	1306	149
21	users.dsic.upv.es/~jsilva/wwv2013/index2.html	197	163	34
22	users.dsic.upv.es/~dinsa/en/index.html	241	74	167
23	www.engadget.com/index.html	1818	768	1050
24	www.bbc.co.uk/news/index.html	2991	364	1360
25	www.vidaextra.com/index.html	2331	1137	1194
26	www.ox.ac.uk/staff/index.html	948	525	410
27	clinicaltrials.gov/ct2/search/index/index.html	543	389	120
28	en.citizendium.org/index.html	1083	414	667
29	www.filmaffinity.com/es/main.html	1333	351	976
30	edition.cnn.com/index.html	3934	192	3742
31	www.lashorasperdidass.com/index.html	1822	553	722
32	labakeryshop.com/index.html	1368	218	962
33	www.felicity.co.uk/index.html	300	232	68
34	www.thelawyer.com/index.html	3349	1293	1580
35	www.us-nails.com	250	184	35
36	www.informatik.uni-trier.de	3085	64	3021
37	www.wayfair.co.uk/index.html	1950	702	437
38	catalog.atsfurniture.com/index.html	340	301	39
39	www.glassesusa.com/index.html	1952	1708	244
40	www.mysmokingshop.co.uk/index2.html	575	407	168

Table 3: Template data of the benchmarks

Id	VL	TT	PT	DT	Notes (peculiarities)
1	9	0	9	0	All pages add a block in the footer that does not belong to the key page.
2	10	10	0	0	
3	4	4	0	0	
4	8	8	0	0	
5	12	5	6	1	The website uses two different footers (hence, two templates). Therefore, some pages only implement partially the template of the key page.
6	26	7	19	0	All pages share the same header and footer but there are pages with a different layout.
7	8	0	8	0	Some pages use two columns while other use three. All of them are different to the key page.
8	24	5	19	0	Some pages use two columns while other use three. All of them are different to the key page.
9	6	6	0	0	
10	6	6	0	0	
11	6	6	0	0	
12	16	16	0	0	
13	10	0	10	0	The submenu appears inside the main content.
14	16	16	0	0	The main content of the key page uses a layout that is different to the other pages.
15	42	42	0	0	
16	13	13	0	0	
17	29	0	29	0	All pages use more blocks than the key page. For instance, advertisement blocks.
18	74	74	0	0	
19	23	23	0	0	
20	77	77	0	0	
21	12	12	0	0	
22	5	5	0	0	
23	65	65	0	0	
24	5	0	5	0	There are several different templates (but they are very similar).
25	7	7	0	0	
26	7	7	0	0	All pages share the same template. There is a breadcrumb inside the main content.
27	36	36	0	0	
28	32	32	0	0	
29	32	32	0	0	
30	13	13	0	0	All pages share the same template, but the header is a bit different between the key page and the other pages.

Table 3: Template data of the benchmarks

Id	VL	TT	PT	DT	Notes (peculiarities)
31	11	11	0	0	All pages share the same template, but the header is a bit different between the key page and the other pages.
32	14	4	0	0	All pages share the same template. There is a big amount of javascript.
33	6	6	0	0	
34	69	69	0	0	
35	10	10	0	0	
36	6	5	0	1	One page linked from the main menu uses a different template.
37	377	377	0	0	
38	6	6	0	0	
39	86	86	0	0	
40	35	35	0	0	

2.3 Guidelines for using the suite

2.3.1 Downloading and configuring the suite

TECO is freely distributed and can be downloaded from the URL:

<http://www.dsic.upv.es/~jsilva/retrieval/teco>

After downloading the suite, a directory that contains 40 folders, one for each website, is created. Table 4 shows the path to the key page of each benchmark.

2.3.2 Rules for using the suite and report

All researchers and developers that use TECO must follow two basic principles:

1. They must publish their results so that they are publicly available.
2. They must provide enough information so that anyone can easily duplicate their experiments.

3 Conclusions

This paper presents a benchmark suite composed of 40 heterogeneous websites. This benchmark suite can be used to test any technique that works with web-pages, but it is specially useful for template detection and content extraction because it includes a gold standard for them. Concretely, the gold standard identifies for each benchmark the template and the main content. Thus, it can be used to evaluate and compare techniques and implementations of these disciplines. The suite is publicly available and free.

Table 4: Path to the key page of each benchmark

Id	Path to the key page
1	pages/water.org/index.html
2	pages/www.jdi.org.za/index.html
3	pages/stackoverflow.com/index.html
4	pages/www.eclipse.org/index.html
5	pages/www.history.com/index.html
6	pages/www.landcoalition.org/index.html
7	pages/es.fifa.com/index.html
8	pages/cordis.europa.eu/fp7/ict/fire.html
9	pages/clotheshor.se/index.html
10	pages/www.emmaclothes.com/index.html
11	pages/www.cleanclothes.org/index.html
12	pages/www.mediamarkt.es/index.html
13	pages/www.ikea.com/gb/en.html
14	pages/www.swimmingpool.com/index.html
15	pages/www.skipallars.cat/en.html
16	pages/www.tennis.com/index.html
17	pages/www.tennischannel.com/index.html
18	pages/www.turfparadise.com/index.html
19	pages/riotimesonline.com/index.html
20	pages/www.beaches.com/index.html
21	pages/users.dsic.upv.es/~jsilva/wwv2013/index2.html
22	pages/users.dsic.upv.es/~dinsa/en/index.html
23	pages/www.engadget.com/index.html
24	pages/www.bbc.co.uk/news/index.html
25	pages/www.vidaextra.com/index.html
26	pages/www.ox.ac.uk/staff/index.html
27	pages/clinicaltrials.gov/ct2/search/index/index.html
28	pages/en.citizendium.org/index.html
29	pages/www.filmaffinity.com/es/main.html
30	pages/edition.cnn.com/index.html
31	pages/www.lashorasperdidias.com/index.html
32	pages/labakeryshop.com/index.html
33	pages/www.felicity.co.uk/index.html
34	pages/www.thelawyer.com/index.html
35	pages/www.us-nails.com/Unternehmen/Ueber_Uns/ueber_uns_l2-dat=5860687c.php.html
36	pages/www.informatik.uni-trier.de/~ley/pers/hd/s/Silva_Josep.html
37	pages/www.wayfair.co.uk/index.html
38	pages/catalog.atsfurniture.com/index.html
39	pages/www.glassesusa.com/index.html
40	pages/www.mysmokingshop.co.uk/index2.html

References

1. Julian Alarte, David Insa, Josep Silva, and Salvador Tamarit. Template Extraction Based on Menu Information. In Josep Silva and Antonio Ravara, editors, *Proceedings of the 9th International Workshop on Automated Specification and Verification of Web Systems (WWV 13)*, page Article 5, 2013.
2. Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a Competition for Cleaning Web Pages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*, pages 638–643. European Language Resources Association, may 2008.
3. David Insa, Josep Silva, and Salvador Tamarit. Using the words/leafs ratio in the DOM tree for content extraction. *The Journal of Logic and Algebraic Programming*, 82(8):311–325, 2013.